知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum
E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【专稿】

网络谣言敏感词库的构建研究

——以新浪微博谣言为例

◎夏松 林荣蓉 刘勘
中南财经政法大学信息与安全工程学院 武汉 430074

摘要：[目的/意义]网络谣言严重影响网络正常信息的传播，对网络谣言进行识别有着重要的现实意义。笔者构建一个基于微博的网络谣言敏感词库，以提高网络谣言的识别精度。[方法/过程]针对微博类社交平台短文本的特点，首先舍弃传统的分词算法，设计 LBCP 抽词算法，并结合位置信息和改进的 TF-IDF 权重来提取敏感词库的种子词集，然后通过聚类算法将种子词的近义词补充到词库中，再将常用的替代词也加入到词库中，从而得到最终的敏感词库。[结果/结论]利用敏感词特征对谣言进行判断，在提取微博的内容特征、用户特征、传播特征以及情感分析特征的基础上，新增敏感词特征以后谣言识别率有明显提升，得到较好的识别效果。

关键词：敏感词库；词向量；特征空间；网络谣言
分类号：G202

引用格式：夏松，林荣蓉，刘勘．网络谣言敏感词库的构建研究：以新浪微博谣言为例[J/OL]. 知识管理论坛，2019, 4(5): 267-275[引用日期]. http://www.kmf.ac.cn/p/185/.

1 引言

对网络谣言进行深入分析有助于及时判断真实或虚假的信息，创造一个健康的网络环境。目前网络谣言的识别多是从用户特征、传播特征的角度进行分析，而事实上，谣言敏感词是识别网络谣言的一个重要特征，谣言敏感词分析有助于提高对谣言的判别，遏制谣言的蔓延和传播。

词库是词汇的集合体，通常包括基本词库以及专业词库，应用较广的专业词库包括流行词库、专业本体词库、敏感词库，情感词库等。其中，现有的敏感词库主要有反动敏感词库、暴恐敏感词库、色情敏感词库、垃圾广告敏感词库等，被广泛地应用在各类贴吧、论坛以及垃圾邮件检测中。

但目前还没有一个完备的网络谣言敏感词库。本文的谣言敏感词库是应用于微博微信这

基金项目：本文系国家社会科学基金资助项目“基于文本挖掘的网络谣言预判研究”（项目编号：14BXW033）研究成果之一。
作者简介：夏松(ORCID：0000-0001-7700-6185)，讲师，博士；林荣蓉(ORCID：0000-0002-5141-6503)，硕士研究生；刘勘(ORCID：0000-0002-9686-9768)，教授，博士，通讯作者，E-mail: liukan@zuel.edu.cn。
收稿日期：2019-06-18 发表日期：2019-09-11 本文责任编辑：刘远颖

类平台的，专门用于谣言的识别，包括：①失实的事件，比如某地发生地震、骚乱等子虚乌有的事件；②夸大事实真相，比如厂商对自身产品的过度或虚假宣传、对同行产品的诋毁；③过期信息的使用及诈骗，比如将小女孩走失的消息更改时间地点或者电话号码之后再次发送，诱导人们拨打有诈骗嫌疑的电话等。这些谣言会在一定时期一定程度上引发了社会各领域人们的关注甚至成为焦点，如果不及时处理，其潜在的安全威胁也是不可估量的，而对于这些谣言涉及的敏感词，传统的词库并不能很好地识别。因此，笔者所构建的敏感词库是基于微博谣言而建立的，有较强的实用价值，为社交平台的谣言识别提供速度和质量的保证。

2 相关工作

敏感词库的构建主要在于敏感词汇信息的识别、敏感词汇的提取以及扩展。其中，敏感信息的提取目前大多通过人工标记与挑选或者基于传统权重计算方法 [1] 去衡量与选择，再基于参考词林，去迭代地识别敏感信息，最后通过相关算法进行敏感词的扩充 [2]，如刘耕等 [3] 采用基于广义的 jaccard 系数方法来计算得到敏感词的相关联词汇。

针对敏感事件和热点话题的很多研究从敏感词库和热点词集入手，取得了较好的效果。词库的构建类似于提取文本中的关键字，多以已有的专业词汇为基础，采用计算特征词权重的方法。徐琳宏 [4] 根据情感分类现状，确定分类的体系，再综合各种情感词汇的资源来构造情感词汇的本体，采用了手工分类以及自动获取结合的方法获取词汇本体；侯丽 [5] 采用 N-Gram 及各种过滤规则结合的术语识别公众日志数据，能较好地识别发现健康类词集；C. Quan 等 [6] 从情感类别符号、情绪强度、情感词、程度词、否定词、连词、修辞等识别情感种子词，从而完成情感词典的构建；F. Peng 等 [7] 利用线性链条件随机场（CRFs）来进行基于字、词、多词等形式的领域集成的中文分词，并通过基于概率的新词检测方法进行新词识别；周强 [8] 提出一种多资源融合自动构建汉语谓词组合范畴语法 (CCG) 词库的方法，不同句法语义分布特征，融合形成 CCG 原型范畴表示，将它们指派给各资源信息完全重合的谓词形成核心词库；K. J. Chen 等 [9] 实现了通过一个未知词提取系统来在线识别新词，主要通过统计信息以及语法语义上下文等信息进行新词识别；彭云等 [10] 在商品情感词的提取过程中，基于商品评论文本，从词义理解、句法分析等角度获得词语间语义关系，并将其嵌入到主题模型，提出基于语义关系约束的主题模型 SRC-LDA，从而实现主题词的提取。

在构造词库时，只是确定了基本词集往往是不够的，需要对其进行扩充从而得到较为完整的词库。词汇扩展与关键字扩展相似，通过词义近似或语义近似展开。H. Chen 等 [11] 从词典中提取了近似语义信息的词作为扩展。S. Yu 等 [12] 利用 VIPS(VIsion-based Page Segmentation) 算法进行查询扩展，该算法主要是通过调用嵌入在 Web 浏览器中的分析器来获取 DOM 结构以及视觉相关信息（所有视觉信息都来自 HTML 元素和属性）进行查询扩展。J. M. Pnote 和 W. B. Croft[13] 提出了将统计语言模型和信息检索相结合，使用词频和文档频率按综合频率对词信息进行排序；T. Pedersen 和 A. Kulkarni[14] 通过聚类实现类似的词的识别，然后将它们应用到语义扩展；P. D. Turney 和 M. L. Litham[15] 通过计算倾向性基准词与目标词汇间相似度的方法识别词汇语义倾向性；A. Neviarouskaya 等 [16] 通过同义词和反义词的关系、上下文语义关系、推导关系以及与已知的词汇单位复合来进行情感词典的扩展。

但是上述敏感词库的构建方法应用于网络谣言语料库建设并不完全合适，首先目前谣言并没有可参考的词林。而且传谣变化形式多样，扩展和传播方式多种多样。某些词汇出现在网络谣言中的频率高，同时存在于正常微博中的频率也高，不能单独用来判定谣言。针对以上谣言敏感词的特点，笔者设计了一个抽词

知识管理论坛，2019(5):267–275
DOI：10.13266/j.issn.2095–5472.2019.028

算法提取敏感词并进行多级扩充，旨在建立一个实用的网络谣言敏感词库。

3 谣言敏感词库设计

3.1 谣言敏感词库构建的困难

针对谣言特征所构建的微博谣言敏感词库是一个专业性偏向性较强的专业词库，需要有大量的微博谣言语料，同时在构建敏感词库的过程中会遇到下述困难：

（1）人工干扰。谣言发布者常常会采取多种方法来逃避关键词的匹配过滤。例如在敏感组合词汇间夹杂了一些无意义的数字与符号，如"抵 @ 制！ 共 &$ 产 &0 党"。然而这类复杂多变的形式，却并不影响人们正常的阅读，这种情况直接进行敏感词库匹配是无法解决的。

（2）准确性。部分在谣言微博中出现的敏感词，很多时候也会出现在正常微博的文本中致使对文本敏感得分的判定很容易出现偏差。即大多词汇只有在特定的语境中才会显示出其谣言的特性。

（3）分词问题。网络用语越来越随意，新词、未登录词等层出不穷以及谣言具有时效性等使得传统的分词工具不适用于此类文本，从而对谣言识别带来影响。

对于第一类人工干扰的谣言文本，即夹杂符号的敏感词的检测与识别，笔者将通过扩充停用词解决，对待测文本分词之后进行去停用词等预处理方法来解决；对于第二类准确性的问题，笔者引入位置权重以及敏感度权重来抽取敏感词，将词汇在谣言与正常微博中的词频比以及位置权重（词汇是否处于标题中）作为衡量的因素，同时对种子词集进行相似词、关联词扩展；对于第三类分词问题，笔者提出基于敏感热度的 L-CPBL 抽词算法摒弃了传统的分词工具，基于内聚度以及外聚度来提取文本片段，以更加适用于网络社交文本。

3.2 总体设计

本研究中网络谣言敏感词库构建的基本思路是：首先收集网络谣言语料，然后利用抽词算法构建种子词集，进而对种子词集进行扩展得到完备的谣言敏感词库，其总体流程如图 1 所示：

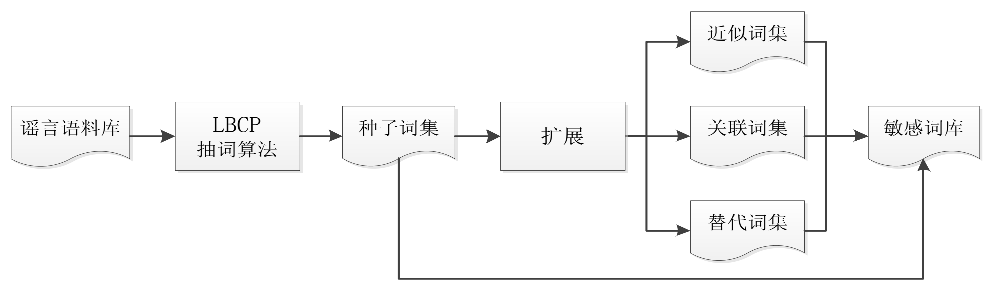


图 1 敏感词库构建流程

因为目前的分词软件大多具有普适性，用来针对某一领域发现特定词、敏感词、新词效果不佳，因此种子词采集没有直接分词，而是设计了 LBCP(Location- Based Cohesion and Polymerization) 算法进行抽词，通过计算词的内聚度和外聚度，结合词权重和位置权重得到种子词集；然后对种子词集从近似词、关联词和替代词等方面进行扩展，最终合并成为谣言敏感词库。

3.3 LBCP 抽词算法

LBCP 抽词算法是考虑了词语位置和上下文信息的抽词过程，其提取谣言种子词集的流程如图 2 所示。

LBCP 抽词算法首先设置一个滑动窗口，从中提取候选词汇，计算候选词汇的内聚度（表示该词的聚合程度）、外聚度（描述该词与上

下文的关联），再考虑该词在文中的位置（标题中取 2，正文中取 1），利用改进的 TF-IDF 权重计算出该候选词的综合得分并进行排序，再提取排在前面的候选词形成种子词集。

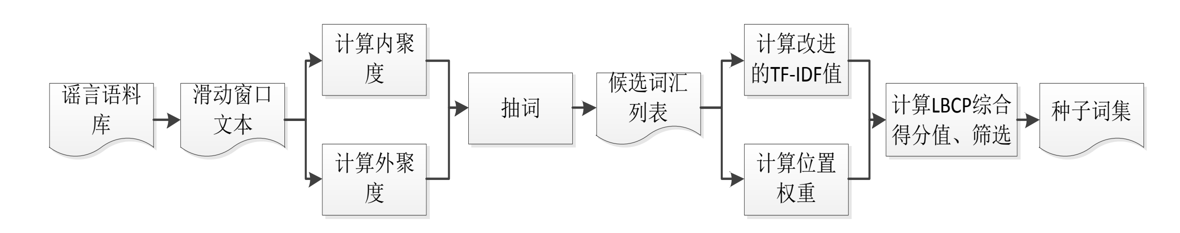


图 2 种子词集提取的流程

网络谣言文本短小随意，新词、流行词汇众多，依赖传统词库容易带来较大误差，笔者采用的方法有效规避了传统分词方法过度依赖词库的问题。利用改进的 TF-IDF 权重计算较好地体现了谣言词汇的信息完整性、词汇的领域相关性，其中的位置权重较好地体现了词汇位置的重要性（词汇是否处于标题中）。

3.3.1 内聚度

内聚度主要用来分词，一般取某一固定长度的窗口，依次滑动窗口找到其中的分词。比如图 3 中的文本，设窗口长度为 6（即每个词不超过 5 个字），则滑动窗口会包含“央视已经报道”6个字，从中计算出“央视”内聚度最高，因此按前 2 个字做划分，提取“央视”这个词。然后滑动窗口右移，取出“已经报道此事”6 个字，进而提取“已经”这个词。依此类推，依次取到文本中的其它词。

具体计算内聚度时，需要先按滑动窗口中的每一个字划分，得到左右两部分，计算两部分在语料库中的出现概率乘积(即 p(左)*p(右))，取最大作为分词的内聚度。例如图 3“央视已经报道”6 个字，依次计算：p(央)*p(视已经报道)、p(央视)*p(已经报道)、p(央视已)*p(经报道)......p(央视已经报)*p(道)，最终 p(央视)*p(已经报道)的乘积最大，因此提取“央视”这个词，并将该乘积作为“央视”的内聚度。

央	视	已	经	报	道	此	事	......
c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	......

图 3 计算内聚度使用的滑动窗口

设滑动窗口中的文本 X 由 n 个汉字 $C_1C_2...C_n$ 构成（见图 3），内聚度 $h(x)$ 的计算首先由公式（1）找到应该划分的候选词位置 i，从而划分出候选词 $x=C_1C_2...C_i$。

$$i = \arg\max\{p(c_1)\cdot p(c_2...c_n),...,p(c_1...c_i)\cdot p(c_{i+1}...c_n),...,p(c_1c_2...c_{n-1})\cdot p(c_n)\} \quad 公式（1）$$

进而利用公式（2）记录词 x 和在本窗口的内聚度 $p_i(x)$，最后利用公式（3）计算该词在所有窗口中的内聚度之和，作为最终的内聚度 h(x)，公式 (3) 中 k 代表词 x 在整篇文档中出现的次数。

$$p_i(x) = p(c_1c_2...c_i)\cdot p(c_{i+1}...c_n) \quad 公式（2）$$

$$h(x) = \sum_{j=1}^{k} p(X_j)/p_j(x) \quad 公式（3）$$

3.3.2 外聚度

单看一个候选词汇的内聚度，可能会出现诸如“的 ...”组合被认为成独立的词。所以笔者还考虑了词的上下文联系，这里用外聚度表示。如果某个词能够被认为是一个独立的词，那么它应该能和各种词搭配出现在不同的语言环境中，即具有丰富的“左集合”和“右集合”。外聚度用左右信息熵来进行衡量。假设词 x 与左边相邻的词汇组成的短语为 x_lx，与右边相邻的词汇组成的短语为 xx_r，则词 x 的外聚度 $g(x)$ 的计算公式如公式（4）所示：

$$g(x) = \min\{-\sum p(x_lx)\log p(x_lx), -\sum p(xx_r)\log p(xx_r)\} \quad 公式(4)$$

因为熵表示不确定性，所以熵越大，不确

知识管理论坛，2019(5):267-275
DOI：10.13266/j.issn.2095-5472.2019.028

定越大，也就是这对词左右搭配越丰富，选择越多。计算一个文本片段的左信息熵和右信息熵，如果该文本片段的信息熵偏大，就可以把它看作一个独立的、可以抽取出来的高频谣言词汇。

3.3.3 改进的 TF-IDF 权重

通过内聚度和外聚度得到基于谣言语料的大量新旧词汇，然后可以利用这些词的权重进行筛选。本文权重基于 TF-IDF 计算，但是做了如下改进：①由于旧新闻重复散播或者同一条谣言仅仅修改了人名、地名、手机号码等张冠李戴型的谣言比较多，因此对文档频率要求较高，而对于逆文档频率要求并不高，不要求词语对文档有独特的标识性，因此对于 TF-IDF 公式中 TF 和 IDF 给予不同的权重（公式（5）中增加 λ，使得 TF 的权重值大于 IDF 的权重值）；②消除了文档长度的不同对词权重的影响（公式（5）中增加分母，进行余弦归一化处理），同时通过对词频取对数来消除词频大小差异对权重计算的影响。词 x 改进的 TF-IDF 权重计算如公式（5）所示：

$$W_1(x)=\frac{(\log_2(\lambda tf+1.0)*\log_2(\frac{N}{(1-\lambda)idf}))}{\sqrt{\sum_i(\log_2(\lambda tf_i+1.0)*\log_2(\frac{N}{(1-\lambda)idf_i}))^2}} \quad 公式（5）$$

其中，N 为微博总条数，λ 是词频 TF 的权重，tf 和 idf 分别表示词 x 的 TF 值和 IDF 值，公式（5）右边的分母利用谣言微博中出现的每个词 x 的 TF 值 tf_i 和 IDF 值 idf_i 进行余弦归一化处理。

3.3.4 位置权重

针对微博谣言来说，微博标题或者话题中的词汇比内容词汇更具有代表性，更能反映出微博的主题，也就更能反映出微博主题的谣言敏感度。对微博文本中词 x 的单个位置权重值定义如公式（6）所示：

$$L_i=\begin{cases}2 & i\in title\\ 1 & i\notin title\end{cases} \quad 公式（6）$$

对微博内容进行扫描，如果有"【】"或者"[]"或者 #，则将其视为微博的标题或者话题，将其单个位置权重值为 2，否则为 1。词 x 位置权重为公式（7）所示：

$$w_2(x)=\frac{\sum_{i=1}^{D}L_i}{D} \quad 公式（7）$$

其中，$w_2(x)$ 为词汇的位置权重。L_i 表示某条微博 i 中词 x 的位置权重值，D 表示包含词 x 的微博总条数。

3.3.5 抽词算法流程

LBCP 抽词算法的步骤如下：

Step 1: 利用公式 (1)、公式 (2)、公式（3）进行分词（词的长度不大于某个阈值 t），并计算各分词 x 的内聚度 h(x)；

Step 2: 利用公式（4）计算各分词 x 的外聚度 $g(x)$，求得每个词的内聚度和外聚度之和，筛选出其和大于某阈值 θ 的词汇；

Step 3: 对筛选出的词计算其改进的 TF-IDF 权重 $w_1(x)$ 和位置权重 $w_2(x)$，根据公式（8）计算综合得分值；

$$LBCP(x)=w_1(x)\times w_2(x) \quad 公式（8）$$

Step 4：按综合得分 $LBCP(x)$ 进行排序，取前 M 个词汇作为种子词集。

3.4 扩展词集

为了实现对谣言敏感词的有效扩展，笔者主要从种子词集的近似词、关联词以及替代词 3 个方面进行词库的扩展。

3.4.1 近似词集

Word2Vec 方法计算的词向量能反映词的上下文和语义关系，因而近似词集的扩展主要通过词向量 Word2Vec 进行计算，再通过聚类找种子词的相似词，从而得到基于上下文和语义关系的近似词，其流程如图 4 所示：

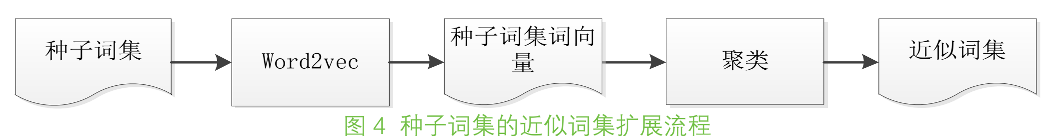


图 4 种子词集的近似词集扩展流程



知识管理论坛
2019 年 第 5 期(总第 23 期)

3.4.2 关联词集

单个敏感种子词可能出现在谣言中，也可能出现在正常微博中，比如“免费”这个词，它既可能出现在不良厂商的微博谣言中，也可能出现在正规商家的微博宣传中，但当“免费”和“转发”共现时，它就极大可能是谣言。因此，对于每个种子词计算其高频率共现的词汇，即与之相关度高的词汇，这些词汇有助于提高谣言的识别率。

笔者采用互信息的方法来寻找种子词关联词集。种子词的关联词集的构造流程如图 5 所示：

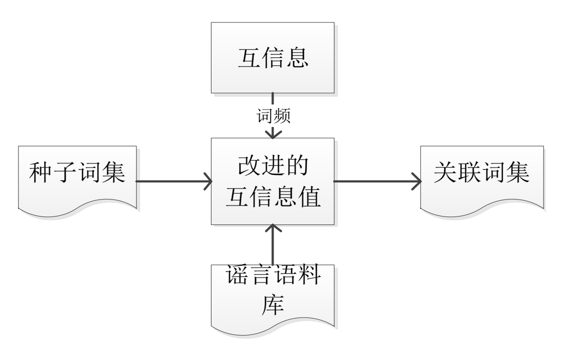


图 5 种子词的关联词集的构造流程

但是谣言中这样产生的成对词互信息较高，词频却较低，这样对于谣言的识别作用不大，因此在互信息计算上加入了词频信息，计算如公式（9）所示：

$$PMI(x,y)=\mathrm{p}(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad 公式(9)$$

其中，$p(x,y)$ 为词 x 和 y 的词频。

3.4.3 替代词集

谣言信息发布者通常会采取多种方式来逃避敏感词匹配过滤，比如把谣言敏感词进行中英文的转换或者缩写等，因此也需要找出种子词的替代词集。这样的词处理量并不多，本研究通过人工来完成，比如：

（1）拼音：拼音代替汉字，如“拐走”——“guai 走”。

（2）英文：英文代替汉字，为种子词的英文翻译。

（3）缩写或简写：种子词的常用缩写形式，如“神州六号”——“神 6”。

经过以上 3 种扩展，最终将种子词集与各扩展词集合并构建了网络谣言的敏感词库。

4 实验

4.1 数据集

实验爬取了新浪微博社区管理中心、“谣言粉碎机”以及各地区辟谣平台上发布的 30 034 条谣言。同时爬取了包括中国新闻网、央视新闻等 35 000 余条正常微博作为正类数据。这些数据都经过了包括去噪声、去停用词等处理过程。去噪主要是删除了总长度不足 5 个字的微博，这类微博多携带信息较少，处理的意义不大，删除后可提高处理效率。

4.2 提取种子词集

利用第 3.2 节中种子词的抽取思路，将微博谣言文本中长度不超过阈值 t（本文取值为 9）的文本都当作潜在的词，通过样例数据实验确定内聚度和外聚度的阈值，最后提取出所有不大于阈值 t 的候选词。在处理过程中，把全部微博谣言语料作为一个整体，利用 LBCP 抽词算法提取候选词 43 363 个。候选词的内聚度、外聚度及其在谣言微博和正常微博中的词频如图 6 所示：

id	infoent	polymerization	count-rumors	count-true
灾情	3.935167502	1.841074735	13	11
造谣	5.433056885	0.517786403	106	3
快报	2.826583357	0.449818199	25	54
常青	2.079441542	0.973234765	3	9
小孩子	6.007970388	2.087595151	39	4
老者	1.464816385	0.241533882	2	1
信号	4.390551062	0.49354779	27	35
回答	5.648174794	0.890389663	66	69
1000	3.487642478	8.816360165	485	61
堂堂	1.464816385	0.183307508	2	2
桐庐	2.144179782	0.362506117	10	6
忽悠	4.233782827	1.143519054	10	17
京制	1.464816385	0.381876902	2	1
幸运	2.609454099	0.38634769	28	12
血肉	1.934479312	0.227439139	12	5
酿成	2.270885977	0.309373755	6	4
坠毁	2.1520804	9.408423209	70	60
县政府	2.857433791	0.216519139	12	3
逃逸	1.87010847	0.365290265	13	32
见到	4.457286529	0.976786252	90	61

图 6 候选词汇的内聚度、外聚度及词频

在上述结果的基础上，结合位置权重因子，根据 LBCP 综合值进行排序，取前 300

知识管理论坛，2019(5):267-275
DOI：10.13266/j.issn.2095-5472.2019.028

个作为谣言种子词集。通过 LBCP 抽词算法挑选出来的部分种子词集如表 1 所示：

表 1 谣言种子词集

类型	词集示例
种子词集	孩子、拐走、扩散、酬金、紧急、严重、知情、女孩、帮忙、转发、求、找、联系、死、爆炸、死亡、转转、钱、伤、黑、白血病、救、去世、农药、丢、癌症、偷、救援、专家、卖、滋、食品、导致、真相、死了、批捕、感染、必须、提醒、杀、禁用、失踪、抢救、证实、罪、打死......

4.3 种子词集的扩展

（1）近似词集扩展。利用 Word2Vec 工具计算得到 300 个种子词集的词向量，再分别计算各个词向量维度上的均值，计算得到种子词集的均值向量。用词集的平均向量利用 KNN 模型聚类，得到 300 个种子词最相近的词。实验中共取到与种子词集同属一类的 1 785 个词作为扩展的近似词，形成的近似词集如表 2 所示：

表 2 近似词集部分示例

类型	词集示例
近似词集	朋友、监控、附近、万分、抱、留意、兄弟、男人、告、姐妹、联系人、双重、关心、婴儿、达、恒天、教授、主、逸、场合、伤亡、死伤、电视、中伤、保、记、妇女、鬼子、保健、院、日本、含、果断、刷、毒素、局、发现、天地、血压、此次、余香、喝、引发、插头、真的、兰、版、海域、七、警、接力、唯一、轻、人数、新闻、消防、天津、牺牲、杭州......

（2）关联词集扩展和替代词集。计算种子词集与语料库中其他词的互信息的大小，并降序排列，得到最终有 175 个词汇的关联词集。如“拐走”的扩展词包括：找到、造谣、逝去、倒霉、最近、诅咒、资助、真相、折磨等。

种子词替代词集包含了 300 个种子词的拼音、英文以及缩写简写形式，如帮忙的替代词有：Help、bangmang、bm 等，酬金的替代词有：Remuneration、fee、pay、choujin、cj 等。

至此，整个谣言敏感词库构建完成，敏感词库包含种子词集 300 个，近似词集 1 785 个，关联词集 175 个，替代词集 300 个，共计 2 260 个。

4.4 微博谣言识别

实验另外爬取了 2018 年 1 月到 2018 年 3 月期间，新浪微博“谣言粉碎机”以及各地区辟谣平台上发布的 5 000 条谣言数据，同时爬取了包括中国新闻网、今日头条、央视新闻在内的微博大 V 账号的正常微博 5 000 条。将这 10 000 条微博作为测试数据，以验证敏感词库对谣言识别的提升作用。

从混合的 10 000 条微博数据中提取传统特征和敏感词特征，将其作为输入数据。传统特征包括发布该微博的用户信息（用户粉丝数、关注数、注册事件、已发布微博数量、是否验证用户）、微博的结构特征（转发数量、微博的长度、是否包含”@”、是否包含标签、是否包含 URL、是否含有表情符号、标点符号的使用情况、是否含有第一人称等）以及每条微博所有词的词向量加和平均值。敏感词特征包括敏感词的个数和敏感词得分总和。

利用以上提取的微博特征，通过随机森林、SVM、GBRT、CNN、BiLSTM、TextCNN 等分类模型构建微博谣言分类器。由于重点在谣言的识别，因此，本文要求谣言的召回率（本身是谣言且被正确识别出来的比例）尽量大，准确率尽量高。实验中采用十折交叉验证，多种算法的准确率和召回率在加入敏感词库特征前后的对比结果如表 3 所示：

表 3 敏感词库特征对谣言判别的效果

判别模型	传统特征		传统特征+敏感词特征	
	准确率	召回率	准确率	召回率
随机森林	79.82%	62.98%	89.29 %	85.10%
GBRT	81.44 %	65.65%	92.65%	86.71%
SVM	80.71 %	66.09%	85.94 %	83.22%
CNN	80.38%	72.66%	91.12 %	83.54%
BiLSTM	82.68 %	73.54%	95.26%	88.67%
TextCNN	81.25 %	77.12%	93.48 %	86.09%

通过实验可知，当敏感词特征和传统特征融合之后，各种分类方法的准确率和召回率都有大幅的提升，其中 BiLSTM 的准确率超过 95%，召回率也接近 90%。可以看出，谣言敏感词库的构建在提升微博谣言的识别率方面达到了预期的效果。

5 结语

网络谣言敏感词库是谣言识别的重要基础，笔者旨在构建敏感词库并用辅助实验证明对谣言微博识别的有效性。利用大量语料库，笔者构建了一个基于敏感热度 L-CPBL 抽词算法及其相似词和扩展词的谣言敏感词库。第一步是种子词集的提取，L-CPBL 抽词算法是一种无词典参考的快速抽词算法，同时结合改进的 LTC 权重以及位置权重因子，对谣言敏感词库的种子词集的提取更准确；然后基于词向量模型空间优化以及聚类算法对种子词集进行扩展，综合得到适用于谣言的敏感词库。笔者构建的敏感词库适用于微博类社交短文本，并且构建过程不依赖于人工专家的识别挑选，可基于语料库同步更新，因此节省了时间与费用，提高了效率。

笔者创建的谣言敏感词库具有时效性，需要不断收集大量谣言语料，而谣言语料需依赖官方公布的谣言信息作为标注语料，使得敏感词库的更新需要消耗较多的时间和资源，可对敏感词库的更新进行进一步研究，引入时序算法或者从传播的方面进行研究，以便更好地解决时效性问题。

参考文献：

[1] 徐建民 , 王金花 , 马伟瑜 . 利用本体关联度改进的 TF-IDF 特征词提取方法 [J]. 情报科学 ,2011, 29(2): 279-283.

[2] 周晓 . 基于互联网的情感词库扩展与优化研究 [D]. 沈阳 : 东北大学 , 2011.

[3] 刘耕 , 方勇 , 刘嘉勇 . 基于关联词和扩展规则的敏感词库设计 [J]. 四川大学学报 (自然科学版), 2009, 46(3): 667-671.

[4] 徐琳宏 , 林鸿飞 , 潘宇 , 等 . 情感词汇本体的构造 [J]. 情报学报 , 2008,27(2): 180-185

[5] 侯丽 , 李姣 , 侯震 , 等 . 基于混合策略的公众健康领域新词识别方法研究 [J]. 图书情报工作 , 2015,59(23):115-123.

[6] QUAN C, REN F. Construction of a blog emotion corpus for Chinese emotional expression analysis[C]// Proceedings of conference on empirical methods in natural language processing. Stroudsburg：Association for Computational Linguistics,2009:1446-1454.

[7] PENG F, FENG F, MCCALLUM A. Chinese segmentation and new word detection using conditional random fields[C]//Proceedings of international conference on computational linguistics. Stroudsburg: Association for Computational Linguistics,2004:562-569.

[8] 周强 . 汉语谓词组合范畴语法词库的自动构建研究 [J]. 中文信息学报 , 2016,30(3): 196-203.

[9] CHEN K J, MA W Y. Unknown word extraction for Chinese documents[C]// Proceedings of international conference on DBLP. Taipei: Morgan Kaufmann Publishers, 2002:169-175.

[10] 彭云 , 万常选 , 江腾蛟 , 等 . 基于语义约束 LDA 的商品特征和情感词提取 [J]. 软件学报 , 2017,28(3):676-693.

[11] CHEN H, LYNCH K, BASU K, et al. Generating, integrating and activating thesauri for concept-based document retrieval[J]. IEEE intelligent systems and their applications, 1993,8(2):25-34.

[12] YU S,CAI D,WEN J,et al. Improving pseudo-relevance feedback in web information retrieval using Web page segmentation[C]//Proceedings of the 12th international conference on World Wide Web. New York: ACM, 2003:11-18.

[13] PNOTE J M,CROFT W B. A language modeling approach to information retrieval[C]//Proceeding of the 21st International ACM SIGIR conference on research and development in information retrieval. New York: ACM, 1998:275-281.

[14] PEDERSEN T, KULKARNI A. Identifying similar words and contexts in natural language with sense clusters[C]// Proceedings of the 20th national conference on artificial intelligence. Pittsburgh: AAAI Press, 2010:1694-1695.

[15] TURNEY P D, LITTMAN M L. Measuring praise and criticism: inference of semantic orientation from association[J]. ACM transactions on information systems,

DOI：10.13266/j.issn.2095−5472.2019.028

2003, 21(4):315-346.

[16] NEVIAROUSKAYA A,PRENDINGER H,ISHIZUKA M. SentiFul: a lexicon for sentiment analysis[J].IEEE transactions on affective computing,2011,2(1):22-36.

作者贡献说明：

夏 松：设计模型，完成实验，修改论文；

林荣蓉：采集数据，进行实验，撰写论文初稿；

刘 勘：提出研究思路，设计研究方案，修改论文与定稿。

Construction of Sensitive Thesaurus for Network Rumors
——Taking the Microblog Rumors as an Example

Xia Song Lin Rongrong Liu Kan

School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430074

Abstract: [Purpose/significance] The network rumors seriously influent the spread of normal information on the internet. The purpose of this paper is to construct a sensitive lexicon on microblog rumors and to improve the recognition accuracy of the network rumors. **[Method/process]** According to the characteristics of microblog's short text on social networking platforms, this paper focuses on construction of the microblog sensitive thesaurus, which is built up through LBCP algorithm and extension of multiple level words. At first, the method directly extracts words through LBCP algorithm, which considers the cohesion and polymerization of rumor words. And then, based on the core words, multiple level words are expanded to get sensitive thesaurus. **[Result/conclusion]** In addition to the features of the text, user characteristics, propagation characteristics, emotional analysis, and rumor features based on sensitive thesaurus are exploited. Experimental results show that the accuracy of microblog's rumor recognition can be improved greatly based on sensitive thesaurus.

Keywords: sensitive thesaurus word embedding feature space network rumors